Policy Analysis in the health-services market: accounting for quality and quantity

Bernard Fortin, Nicolas Jacquemet, Bruce Shearer

Policy Analysis in the health-services market: accounting for quality and quantity *

Bernard Fortin†, Nicolas Jacquemet‡, Bruce Shearer†

March 2008
CIRPEE DP n° 08-07

Abstract

We provide a theoretical and empirical framework for evaluating the effects of policy reforms on physician labor supply. We argue that any policy evaluation must account for both the quality and the quantity of services provided. The introduction of quality into the analysis has implications for both the theoretical and empirical analysis of labor supply, and consequently policy evaluation. In particular, endogenous quality choices introduce non-linearities into the budget constraint since the marginal return to an hour of work depends on the quality of services provided. We illustrate by considering a particular example: the recent reform in compensation contracts for specialist physicians in the province of Quebec (Canada). Prior to 1999, most Quebec specialist physicians were paid fee-for-service contracts; they received a piece rate for each clinical service provided. In 1999, the government introduced a mixed remuneration system, under which physicians received a base (half-daily or daily) wage, independent of services provided, and a reduced fee-for-service. Moreover, the government allowed physicians to choose their contract. We derive theoretical results for the effect of the reform on the quantity and quality of services supplied by analyzing "local" prices and virtual income. We propose discretizing the choice set as an empirical approach to policy evaluation in the presence of non-linear budget constraints.

Keywords: Health production, Quality of health services, Discretized models.

JEL codes: I12, D11, C25.

*We wish to thank two anonymous referees for useful comments. Financial support from both the Canada Research Chair in Social Policies and Human Resources and the Canadian Institute of Health Research are gratefully acknowledged.

†Laval University (CIRPEE) and CIRANO; ✆Bernard.Fortin@Ecn.Ulaval.Ca ; ✉Université Laval–Département d'économique, Ste-Foy (Qc), G1K 7P4 Canada.

‡University Paris 1, Panthéon–Sorbonne and Paris School of Economics ; ✆Nicolas.Jacquemet@univ-paris1.Fr; ✉Centre d'Economie de la Sorbonne, 106 Bd. de l'hôpital, 75013 Paris, France.

1 Introduction

In countries where health care is provided in the public sector, contractual design is an important policy tool. Economists have studied the ability of incentives to induce physicians to locate in remote areas (Bolduc, Fortin, and Fournier, 1996) and to limit medical expenditures (Gaynor, Rebitzer, and Taylor, 2004). Yet, while compensation policies affect the labor supply of physicians, standard labor supply models are generally ill equipped to evaluate the social benefits of different contracts. This is partly due to the multitasking nature of the supply of health services. Indeed, physicians can adjust both the quantity and quality of services, both with consequences for social welfare.[1]

Analyzing policies within this setting requires modeling physician decisions with respect to both the quantity and quality of services provided. However, introducing quality (as proxied by the average time spent per service) into labor supply models leads to non-linearities in the budget constraint. For example, the opportunity cost of leisure cannot be taken as constant if the value of an hour of work depends on (endogenous) quality. These non-linearities complicate both the theoretical and the empirical analysis of the labor supply, and hence policy evaluation. In this paper we present a framework for evaluating policies in the presence of these non-linearities. We develop a theoretical model of physicians' choices regarding the quantity and quality of services provided. We show that the model can be used to analyze policy (contractual) changes by considering the income and substitution effects that those changes induce. Theoretical results are derived by considering a linearization technique first developed in the literature on fertility (Becker and Lewis, 1973). We also show that it is possible to empirically evaluate different policies within non-linear setting by estimating discrete-choice labour-supply models (van Soest, 1995).[2]

While the approach is generally applicable, we present our results within the context of a particular example: the recent reform of specialist physicians contracts in the province of Quebec (Canada). Prior to 1999, most Quebec specialist physicians were paid fee-for-service (FFS) contracts. Under these contracts, a physician received a piece rate for each service provided. In 1999, the Quebec government introduced a mixed-remuneration (MR) contract, giving physicians a base (half-daily or daily) wage, independent of the number of services provided, and a reduced (vis à vis the FFS contract) piece rate for services provided. Importantly, there was a voluntary aspect to the reform. Once the government made the MR contract available, different physicians could select the contract under which they preferred to work.

[1] Recent recognition of the importance of quality has led some governments to introduce quality bonuses into physician pay; see, for example, Carroll (2003) and Rosenthal, Fernandopulle, Song, and Landon (2004) for the US case. The most comprehensive quality-bonuses experiment is the introduction in 2002 in the UK of a quality score on which up to 18% of physicians' remuneration is based (Shekelle, 2003; Smith and York, 2004). The cost of such an elaborated performance pay is the complexity of the system, wich seems to be responsible for the relative failure of the reform (Smith, 2003).

[2] A more detailed econometric analysis as well as empirical results derived from the approach we propose here are provided in Fortin, Jacquemet, and Shearer (2006).

The basic goals of the government in introducing the MR system were threefold. First, the government sought to control costs since the MR system reduces the fee-for-service and therefore the marginal gain to seeing patients unnecessarily. Second, it was hoped that the reform would induce physicians to increase the amount of time devoted to teaching and administrative duties by reducing the monetary penalty imposed on physicians who perform these duties under FFS. Also, the reform intends to reduce inequities between physicians devoting much of their time on these duties and the others. Third, the government sought to improve the quality of health care by increasing the amount of time physicians spend to each patient.

We provide a detailed theoretical analysis of the impact of voluntary switching from FFS to MR on labor supply, with emphasis placed on the impact of the reform on the quantity and quality of health services. Theoretical results are derived by linearizing the physician's budget constraint at the optimum and considering "local" (or shadow) prices and virtual income (Blomquist, 1989). We derive conditions under which the income and substitution effects can be signed. Perhaps unsurprisingly, the theoretical effects of the reform are generally ambiguous, suggesting an important role for empirical work. Discrete choice labor supply models have two main benefits in this context. First, they allow for the estimation of structural parameters in a flexible statistical framework, permitting the evaluation of existing and potential future policies. Within the context of the Quebec reform, the voluntary nature of the reform implies that the labor supply effects observed on those physicians opting for the MR contract are unlikely to be representative of a mandatory policy. Second, the non linearities in the budget set implied by the endogenous quality choices are handled in a quite straightforward and flexible manner.

The rest of the paper is organized as follows. In the next section we provide a brief overview of recent developments in the health economics literature, highlighting the trade-off between the quality and quantity of services provided. We also describe the features of the incentives schemes that are the focus of this paper. The theoretical analysis is provided in Section 3, where we first introduce the basic assumptions of the model (Section 3.1) and then develop our methodology (Section 3.2). We provide sufficient conditions for a positive response of labor supply to incentives in Section 4.2. In Section 5, we outline the steps needed to estimate a discretized model of physicians' preferences. The results and contributions of the paper are summarized in Section 6.

2 Physicians incentives

This section summarizes the main existing results regarding physicians' practice response to incentives.[3] As a first step, it is necessary to specify the type of economic agent that best captures physician behavior. In fact, the theoretical literature has alternatively modeled physicians as either a firm or as a supplier of labour.[4] As stressed by McGuire and Pauly (1991), a model with profit maximization unambiguously predicts a reduction in services as a result of a reduction in

[3] See, *e.g.*, Aas (1995) for an earlier survey.

[4] McGuire (2000) provides an extensive discussion of physicians motives.

fees paid for services. Alternatively, Rice (1983) shows that if physicians behave so as to maintain income at a desired level, then a reduction in fees will lead to an increase in services delivered. Utility maximization represents an encompassing assumption regarding physician motives, since it reduces to either profit maximization or target income depending upon the relative size of the income and substitution effects (McGuire and Pauly, 1991). For the remainder of this analysis we will assume that physicians maximize utility.

2.1 Piece rates: promoting productivity

Under a piece rate scheme, each service (or, possibly, patient[5]) is remunerated at price P. The income earned by a physician providing A services during the chosen unit of time is hence $X = P\,A$. Such a fee-for-service (hereafter FFS) contract links the income earned to productivity, as measured by the amount of services delivered. This type of contract has been shown to increase the productivity of manual laborers vis à vis fixed wages (Lazear, 2000; Paarsch and Shearer, 1999; Shearer, 2004) and has been confirmed in the case of physicians (Hemenway, Killen, Cashman, Parks, and Bicknell, 1990). Piece rates are however known to be problematic when available performance measures imperfectly reflect the goals of the principal (Baker, 1992). This is particularly problematic regarding the compensation of physicians, due to information asymmetries and the high variety of tasks that have to be completed.

A physician's practice involves information asymmetries with both the patient and the principal (the government in a public-health care system). Regarding patients, physicians hold private information about the health status of the patient and hence on what kind of services should be delivered (Arrow, 1963). As a result, the demand for their services may be endogenous to physicians' behavior (Evans, Parish, and Sully, 1973); physicians may be able to induce their own demand by prescribing services beyond what health requires. This "demand inducement" implies that the productivity measure can be used by physicians as an instrument to control their income.[6] With the notable exception of Norwegian physicians (Carlsen and Grytten, 1998, 2000; Sørensen and Jostein, 1999; Grytten and Sørensen, 2001), empirical studies have found evidence for demand inducement in a wide range of developed countries, including France (Delattre and Dormont, 2003), the US (Gruber and Owings, 1996; Nguyen and Derrick, 1997) and Canada (Schaafsma, 1994), as well as in the province of Quebec (Rochaix, 1993; Nassiri and Rochaix, 2006). Given this, one would expect demand inducement to be undermined by breaking the link between income and the number of services delivered (Grytten and Sørensen, 2001) through,

[5]Capitation is not considered here. See Selden (1990) for a theoretical analysis, Hutchison, Birch, Hurley, Lomas, and Stratford-Devai (1996) and Gosden, Forland, Kristiansen, Sutton, Leese, Giuffrida, Sergison, and Pedersen (2001) for empirical assessments.

[6]A large theoretical literature has addressed the legitimacy of this induced-demand problem. The most famous controversies opposed Feldman and Sloan (1988) to Rice and Labelle (1989) as well as Labelle, Stoddart, and Rice (1994a,b) and Culyer and Evans (1996) to Pauly (1994b,a). See De Jaegher and Jegers (2000) for a recent contribution based on physicians' optimization behavior. The persistent empirical evidence supporting this phenomenon appears to have resolved this debate. Demand inducement is now widely recognized as potentially problematic when using services to measure productivity Fuchs (1986).

e.g., a fixed salary.

Beyond the services delivered, the rate at which those services are remunerated can also be manipulated by physicians. The reason is that physicians hold private information regarding the services a patient received. The bill serving as the basis of the piece rate remuneration is then left under the physicians' control. Income can therefore be adjusted by misreporting the nature of the services supplied. The incentives for this "billing creep" are increasing in the spread between the rates remunerating different services (Evans, 1983). Moving away from those high powered incentives that closely reflect the nature of the task (difficulty, time spent, ...) is then required to contain such a behavior.

Imposing bounds on the relative piece rates that can be implemented seems particularly problematic due to the variety of tasks a physician is expected to accomplish. In fact, the reliability of the diagnosis, the management of health care establishments, as well as the quantity, quality, and suitability of services are among the many tasks included in supplying health services. In such a multitasking environment, it is well known that incentives will fail to guide the agent's (physician's) behavior unless the remuneration scale closely reflects the goals of the principal (Holmstrom and Milgrom, 1991). What is more, incentives are offered only for those tasks that can be measured, such as the number of services. This naturally leads physicians to concentrate on those tasks that are paid piece rates. Since quality is difficult to measure (Stiglitz, 1975), incentives are often implemented at the cost of reduced quality; see Paarsch and Shearer (2000) for an empirical investigation of this issue.

2.2 Fixed wages: Improving quality

Fixed wages represent an extreme departure from high powered incentives. Under this scheme, a physician earns a fixed amount W provided he supplies a minimum hours requirement (typically, the hours spent at work must be above a threshold $\overline{h}$). As a result, the physicians are free to diversify their practice (by spending time to tasks such as teaching duties, management of health care establishments and more generally to "non-clinical hours" or "without patient" hours of work). What is more, the investment in quality is rendered costless from the physician's point of view. This is however obtained at the cost of lower productivity.

Ferrall, Gregory, and Tholl (1998) used Canadian data on physicians' hours of work to show that fixed wages tend to induce a more diversified practice. More precisely, they observe that physicians practicing under fixed wages spend, on average, 5.5 more hours a week at work than their FFS counterparts, while devoting 5.9 hours less to seeing patients. Although this result suggests that more effort is supplied to activities that do not involve seeing patients, the decrease in the time devoted to patients could imply either an increase or a decrease in the quality of services provided, depending on the associated quantity. Time spent on each single service is often taken as a measure of quality (Glazer and McGuire, 1993). Increases in time spent allow more careful examinations, better communication with the patient, as well as more thorough medical testing. In this respect, fixed wages do seem to improve the quality of services; in a

survey of the empirical literature Gosden, Pedersen, and Torgerson (1999) conclude that services delivered by fixed-wage physicians are more lengthy and that each patient receives fewer services. This suggests that paying a fixed wage, not only improves quality, but also contains demand inducement. Similar results have been found when studying the prescription patterns of fixed-wage and FFS physicians. Epstein (1986) estimates that fixed wage physicians prescribe 50% fewer services than FFS physicians. Hemenway, Killen, Cashman, Parks, and Bicknell (1990) document a 23% increase in the number of laboratory tests prescribed by physicians switching from a fixed wage to an FFS scheme.

On the cost side, this decrease in induced demand can be expected to make health care less expensive, while keeping the wealth of the population constant. These improvements in the quality of medical practice are, however, accompanied by a decrease in productivity, since fixed wages induce a lower production of health services through, *e.g.*, a decrease in the physician effort (Gaynor and Gertler, 1995). The net effect on the cost of the health care system is therefore ambiguous.

2.3 Compensation mixing: containing the cost of quality?

Compensation schemes give rise to a trade-off between productivity – the amount of services obtained at a given cost – and quality, measured as both the time devoted to services and hours diversification. As stressed by Ma and McGuire (1997), incentive schemes that use a single instrument (such as piece rates or fixed wages) cannot obtain two objectives: quality and quantity. This has motivated a growing interest in remuneration schemes based on the mixing of various instruments. One possibility is to combine a fixed wage with the piece rate, giving physician income of $P\,A+W$ over the period. We call this scheme a mixed-compensation system.

A mixed-compensation system can combine the advantages of both the fixed-wage and the piece-rate systems. Under the assumption of excess demand for health care (which seems to accurately describe the market for health care in most industrialized countries), the theoretical results of Ma (1994) and Rogerson (1994) establish that efficiency requires a linear combination between prospective reimbursement – a provisional budget, independent of actual services delivered – and cost reimbursement. The importance of informational asymmetries, however, cannot be ignored; the variable part of the remuneration maintains a correlation between pay and productivity, providing incentives. As shown by Levaggi and Rochaix (2003), however, mixing piece rates with a fixed wage can mitigate the incentives to induce demand.

3 Theoretical Model: Quantity and Quality Decisions in Physician Labor Supply

In this section, we present a theoretical analysis of physician behavior under incentive contracts. We concentrate on the quality-quantity tradeoff in the supply of physician services, allowing physicians to choose the number of services provided as well as the treatment rate (which affects

quality). This gives rise to non-linear budget constraints since the opportunity cost of leisure depends on the treatment rate. We use our model to analyze the response of physicians to the introduction of the mixed compensation system in Quebec in 1999.

3.1 Context: basic assumptions of the model

We consider physician choices over the cost and quality of health care. These choices can be viewed as the inputs of a health-care production process. The remuneration system links this input combination to the income earned by the physician. This, along with a specification of physician preferences, fully describes the physician's optimization problem.

3.1.1 The health production process

Our model is developed assuming daily choices of practice variables, including hours of work and services rendered. Hours of work include both clinical hours, devoted to delivering services, and denoted h^c, and non-clinical hours, devoted to activities such as teaching and administrative management, and denoted h^o. The overall time constraint of a physician is therefore $T = h^o + l + h^c$, where l denotes "pure" leisure and T, the available hours. We assume that physicians can freely allocate their time between both work and leisure and, within work, between clinical and non-clinical hours of work.[7]

We let A denote the total number of services provided by a physician, the variable on which the piece-rate payment is conditioned. We model the quality of services as being negatively affected by the treatment rate, τ, equal to the number of services provided per clinical hour worked. Physicians who spend more time per service can make more careful diagnoses, better explain the necessary treatment and provide more advice to patients concerning health-preserving behavior.

Those two variables – number of services and the treatment rate – summarize the two basic dimensions of clinical practice. The number of services measures the quantity of health provided to the population – the higher the number of services, the healthier the patients treated. The treatment rate measures the (inverse) quality of a given quantity of services – the more time devoted to each service, the better the service. This feature is summarized by the output function of physicians' clinical time: the level of health the practice gives rise to: $s = s(A, \tau)$. The health function hence measures the overall quality of clinical time. We assume s to be increasing in both services delivered and time devoted to each service; *i.e.* $s = s(\underset{+}{A}, \underset{-}{\tau})$.

3.1.2 Physician Utility

We assume physicians maximize utility, specified as a function of consumption, leisure and health. We further assume that non-clinical activities increase utility. This implies that some level of these activities will be supplied even under high-incentive contracts, consistent with recent

[7] This assumption seems reasonable in countries where health care is provided in the public sector, typically characterized by excess demand.

empirical evidence.[8] There are two ways of interpreting this assumption.[9] First, performing teaching tasks may increase influence and prestige. In this view, non-clinical hours of work represent a form of "on-the-job" leisure. Second, it could also be the case that non-clinical hours of work are complementary to clinical hours in the clinical production function (for instance, a physician may have to perform a minimal level of administrative tasks in order to properly treat his/her patients or to access some health-care establishment's equipment). In both cases, this assumption implies that physicians derive some well-being from providing non-clinical hours of work. We keep things general by defining utility over two types of leisure in the model: pure leisure (l) and non-clinical hours of work (h^o), also called hereafter "on-the-job" leisure. Services provided affect utility through their impact on the population's health, s (Dranove, 1988; Rochaix, 1989). This can be interpreted as a kind of ethical concern underlying physicians behavior (see Arrow, 1963 and Evans, 1974). Preferences are represented by the static, one-period utility function: $U = U(\underset{+}{X}, \underset{+}{l}, \underset{+}{h^o}, \underset{+}{s})$.[10]

The constrained maximization of this objective function allows us to analyze the impact of switching to MR on practice patterns. Due to the voluntary nature of MR, a physician will choose to switch to MR only when his maximum utility level is higher under MR than under FFS. In this section, we will assume that this condition is satisfied. This will allow us to focus on the effect of switching to MR on practice variables, *i.e.* h^c, h^o, τ and A.

3.2 Labor supply Choices

Let us first derive the optimization program, conditional on a contract, C.

$$\begin{array}{lll} \underset{\{X,l,h^o,A,\tau\}}{Max} & U = U\left(X, l, h^o, s\left(A, \tau\right)\right) & \\ s.c. & (i) \quad T = h^o + l + h^c & \\ & (ii) \quad A = \tau h^c & \\ & (iii) \quad X = f_C A + y_C & \end{array} \tag{1}$$

In the program (1), equation (iii) represents the physician's budget constraint under contract C. The piece rate is f_C and y_C is the non labor income earned under C. Substituting (ii) and $h^c = T - h^o - l$ (from (i)) into the utility function, the latter can be rewritten as $U\left(X, l, h^o, s\left((T - h^o - l)\tau, \tau\right)\right) = \widetilde{U}\left(X, l, h^o, \tau\right)$. The latter is not necessarily strictly concave, since it depends on the properties of the "reduced" health function $\widetilde{s}(l, h^o, \tau) = s\left((T - h^o - l)\tau, \tau\right)$. However, under the assumption that this function is strictly concave in its arguments, it is easy to show that the reduced utility function will be also strictly concave.[11] This gives rise to the

[8] Fortin, Jacquemet, and Shearer (2006) document that such activities are positively supplied under fee-for-service contracts, inspite of the fact that they are not remunerated. From 1996 to 1999, the average number of non clinical hours of work supplied by FFS physicians in Quebec is around 7.5 hours a week.

[9] We thank an anonymous referee for pointing out the alternative interpretation in terms of complementarities.

[10] The utility function is assumed to be twice continuously differentiable. Moreover, without loss of generality, it is assumed to be strictly concave in its arguments.

[11] Note that one does not have to make these concavity assumptions when estimating the model using a discrete choice approach (see van Soest, 1995).

"reduced" optimization program which we will use for our analysis:

$$\begin{array}{ll} \underset{\{X,l,h^o,\tau\}}{Max} & \widetilde{U}\left(X,l,h^o,\tau\right) \\ s.c. & X - f_C\tau\left[T - h^o - l\right] = y_C \end{array} \tag{2}$$

This budget constraint is non-linear due to the interaction between the quantity of services (implicitly included here in $T - h^o - l$) and their quality $(1/\tau)$, and results in endogenous prices. For example, the marginal return to clinical hours of work $(f\tau)$ depends on the chosen treatment rate. These nonlinearities have implications for both the analysis of labor supply (theoretical and empirical) and for the evaluation of policies. Note that interior solutions to the optimization program (2) imply that (at the optimum) the partial derivatives of the reduced utility function with respect to the arguments are: $U = \widetilde{U}(\underset{+}{X}, \underset{+}{l}, \underset{+}{h^o}, \underset{-}{\tau})$. Note also the budget constraint is nonconvex and therefor can cause multiple solutions of the utility maximization problem. Following Blomquist (1989), we rule all all *complicating* non-convexities.

Empirical evaluations of policy changes that concentrate solely on changes in quantity A, will misrepresent the effect of the reform by the factor $\Delta\tau$, the change in quality of services induced by the reform. In what follows we show that, as in the presence of linear budget constraints, changes in quality (and other elements of labor supply) can be decomposed into income and substitution effects.

3.2.1 A Slutsky Decomposition

We begin by deriving a Slutsky decomposition of the impact of a contractual change on practice behavior. The first order conditions of the reduced program (2) can be solved for the demand for each practice variable, denoted β_j, where $\beta_j \in \{X, l, h^o, \tau\}$, as a function of price and income: $\beta_j = \beta_j(f_C, y_C)$. The optimal levels of clinical hours and services can thus be written: $h^c = T - l(f_C, y_C) - h^o(f_C, y_C)$ and $A(f_C, y_C) = h^c(f_C, y_C)\tau(f_C, y_C)$. These demand functions define optimal choices under the contract C which, in turn, allows to calculate the indirect utility obtained under this contract: $\widetilde{U}_C(f_C, y_C) = \widetilde{U}\left(X(f_C, y_C), l(f_C, y_C), h^o(f_C, y_C), \tau(f_C, y_C)\right)$.

Lemma 1 *The impact of a change in contract (from C_i to C_k) on a given practice variable β_j can be approximated as the result of a substitution effect and an income effect–through the change in utility induced by the contract change, $\Delta\widetilde{U} = \widetilde{U}_{C_i} - \widetilde{U}_{C_k}$. This is summarized in equation (3), where E_U denotes the partial derivative of the expenditure function, $E(f, U)$ (equal to y), $\frac{\partial\widetilde{\beta}_j}{\partial f}$ is the partial derivative of the Hicksian demand for β_j $(= \widetilde{\beta}_j(f, \widetilde{U}))$, and $\frac{\partial\beta_j}{\partial f}$ is the partial derivative of the Marshallian demand for β_j $(= \beta_j(f, y))$. All partial derivatives are evaluated at the optimum of contract C_i.*

$$\Delta\beta_j \approx \frac{\partial\widetilde{\beta}_j}{\partial f}\Delta f + \frac{\partial\beta_j}{\partial y}E_U\Delta\widetilde{U} \tag{3}$$

Proof The impact of a change in contract on optimal behavior is given by :

$$\Delta\beta_j = \beta_j(f_{C_k}, y_{C_k}) - \beta_j(f_{C_i}, y_{C_i})$$

For any compensation scheme C, the expenditure function E is given by : $E(f_C, \widetilde{U}_C)$, where $\widetilde{U}_C$ represents the level of utility reached at the optimum under C. As a result, the optimal β_j is given by $\beta_j = \beta_j(f_C, E(f_C, \widetilde{U}_C))$ and the impact of a change in contract can be approximated by :

$$\Delta\beta_j \approx \frac{\partial\beta_j}{\partial f}\Delta f + \frac{\partial\beta_j}{\partial y}(E_f\Delta f + E_U\Delta\widetilde{U}) = \left[\frac{\partial\beta_j}{\partial f} + \frac{\partial\beta_j}{\partial y}E_f\right]\Delta f + \frac{\partial\beta_j}{\partial y}E_U\Delta\widetilde{U} \tag{4}$$

As established by Blomquist (1989), Shephard's Lemma and Slutsky decompositions can be adapted to the case of non-linear budget sets. Adopting Blomquist's notation, let $g = X - f.\tau\left[T - h^o - l\right]$ stand for the budget constraint faced by a physician. Shephard's lemma in the non-linear case implies that : $\frac{\partial E}{\partial f} = \frac{\partial g}{\partial f} = g'_f$. Defining the Hicksian demand functions for the β_j's, denoted $\widetilde{\beta}_j(f, U)$, as the solution to the expenditure minimization problem for a given utility level $\widetilde{U}$, the Slutsky decomposition is given by: $\frac{\partial\beta_j}{\partial f} = \frac{\partial\widetilde{\beta}_j}{\partial f} - \frac{\partial\beta_j}{\partial y}.g'_f$. These properties together imply: $\frac{\partial\beta_j}{\partial f} + \frac{\partial\beta_j}{\partial y}.E_f = \frac{\partial\widetilde{\beta}_j}{\partial f}$. Substitution into (4) gives the result. ∎

The income and substitution effects, induced by a change in contract, will generally be ambiguous. Consequently, ignoring quality in the measurement of treatment effects can lead to an overestimate or an underestimate on the social benefits of the policy. In the next section, we apply these methods to a specific example: the introduction of a mixed-remuneration system in the province of Quebec.

4 A specific example: Mixed Remuneration

The mixed remuneration scheme introduced in Quebec in 1999 offers two distinct contracts to specialist physicians: the traditional contract is a FFS contract that pays a price P for each service provided.[12] The alternative contract is an MR contract which pays a fixed wage, called a *per diem*, and a reduced (relative to the FFS contract) fee for services provided. The *per diem*, denoted D, is earned provided the number of hours spent at work, h, satisfies: $h > \overline{h}$.[13] The *per diem* remunerates, not only the time spent seeing patients, but also the time spent on administrative duties. Services provided under MR are remunerated at a discounted piece rate, the discount factor being α, $\alpha \in [0, 1]$.[14]

[12] In Quebec, the fees are exogenously fixed by health care administrators. We hence assume non-market prices for services in the model.

[13] The analysis is developed assuming this condition is always satisfied by the hours chosen under MR. The results are unaffected by this assumption due to the voluntary nature of MR: it cannot be rational for a physician to switch to MR if the anticipated hours at work under this scheme are lower than $\overline{h}$.

[14] When introduced, in 1999, the reform implemented a fixed-wage set equal to $D = 300$ CAD paid for each $\overline{h} = 3.5$ hours of work. The discount rate α is highly variable depending on specialty, service and health care establishment of provision. It overall ranges from the 0 to 1, the average being around 0.5.

4.1 Analysis of the reform

4.1.1 The Budget constraint

Under the government policy, each physician could choose their compensation system. Let C_i denote the compensation scheme chosen by physician i, $C_i \in \{FFS, MR\}$. As earlier, the budget constraint a physician faces is given by $X = f_C A + y_C$, but now f_C and y_C can be summarized as:

$$f_C = \begin{cases} P & \text{under FFS} \\ \alpha P & \text{under MR} \end{cases} \quad \text{and} \quad y_C = \begin{cases} 0 & \text{under FFS} \\ D & \text{under MR} \end{cases}$$

The voluntary nature of this reform permits somewhat more precision in characterizing the effects of the reform on labor supply.

4.1.2 Income effects

Our model of physicians preferences (Section 3) imposes several constraints on the shape of income effects. First, the health function is assumed to be decreasing in the treatment rate, due to the induced fall in the quality of services delivered. Since ethical motives underlie a positive link between physician's satisfaction and the health function, the treatment rate appears as a "bad" for physicians. Second, the literature on labor supply traditionally concludes that leisure is a normal good (see, *e.g.*, Pencavel, 1986). In our case, leisure is represented, not only by the traditional "pure leisure" referring to the time enjoyed outside work, but also by "on-the-job leisure" represented by non-clinical hours of work. We assume both types of leisure to be normal goods.

Lemma 2 *The sign of the income effect depends on the sign of* $\frac{\partial \beta_j}{\partial y}$. *Moreover*

1. *If the quality of treatment (or, equivalently, minus the treatment rate) is a normal good (*i.e., $\partial\tau/\partial y < 0$*) then a sufficient condition for a negative response of the treatment rate to switching to MR is* $\frac{\partial \widetilde{\tau}}{\partial f} > 0$.

2. *If both kinds of leisure are normal goods then* $\frac{\partial \widetilde{l}}{\partial f} < 0, \frac{\partial \widetilde{h}^o}{\partial f} < 0$ *are sufficient conditions for an increase in leisure upon switching to MR.*

3. *If both kinds of leisure are normal goods then* $\frac{\partial \widetilde{h}^c}{\partial f} > 0$ *is a sufficient condition for clinical hours to decrease upon switching to MR.*

Proof The proof follows from applying (3) to the *voluntary* adoption of MR. By definition, the expenditure function is such that: $E_U > 0$. Moreover, since physicians are free to adopt MR, a physician choosing the new compensation scheme therefore satisfies: $\Delta U|_{MR} \geq 0$; the sign of the income effect depends only on $\frac{\partial \beta_j}{\partial y}$. Last, note that the price variation is negative since $\Delta f|_{MR} = (1-\alpha)P - P = -\alpha P \leq 0$.

1. If the quality of treatment (minus the treatment rate) is a normal good, the income effect on the treatment rate resulting from a voluntary switch to MR is negative. Given the price variation is negative, the result follows directly.

2. If both leisures are normal goods, the income effect on leisure resulting from a voluntary switch to MR is positive. The result follows immediately from the fact that the price variation is negative.

3. Due to the time allocation constraint (ii) in (1), the response of clinical hours of work can be deduced from: $\frac{\partial h^c}{\partial y} = \frac{\partial}{\partial y}\left[T - l - h^o\right] = -\frac{\partial l}{\partial y} - \frac{\partial h^o}{\partial y} < 0$. Therefore, clinical hours of work decrease upon switching to MR if $\frac{\partial \tilde{h}^c}{\partial f} > 0$.

Lemma 2 shows that under certain (reasonable) conditions we can sign the income effect on practice variables arising from a voluntary switch in compensation systems. The overall effect depends on the the sign of the substitution effect. In the following sections we show that signing the income effect is not sufficient for signing the treatment effect – the substitution effects of a contractual change are generally ambiguous.

4.1.3 Price effects

We now provide a framework for the characterization of price effects induced by MR. In our model, physicians simultaneously choose, not only the level of their clinical hours of work (or, equivalently, their gross leisure), but also the treatment rate. The prices appearing in the budget constraint are therefore endogenous. For example, the marginal return to an hour of work depends on the treatment rate. This leads to non-linearities in the budget constraint and adds complexity to the comparative statics.

To analyze price effects in this setting, we rewrite the budget constraint as a linear function of local prices.[15] This technique consists of defining a transformed optimization program, linear in local prices and in the virtual income. Let π_{β_j}, $\beta_j \in \{l, h^o, \tau\}$, denote the local price of each practice variable of interest. Each local price is defined as the marginal price at the optimum. The virtual income is computed as the total expenses on these three goods at the optimum using the marginal prices. Therefore the optimal demands derived from the original program (2) are obtained from the transformed program.

In classic demand theory with linear budget constraints, the optimal demand for a good x satisfies $MRS_{x,x_0} = p_x$, where x_0 is the numeraire. The local prices must be such that the optimal demands derived from this condition are identical to the solutions to (2). It then stems from a simple derivation of (2) that:

$$\pi_l = MRS_{l,X} = f_C\tau \ ; \ \pi_h^o = MRS_{h^o,X} = f_C\tau \ ; \ \pi_\tau = MRS_{\tau,X} = -f_C h^c \tag{5}$$

The program (2), describing the behavior of physicians, is therefore formally equivalent to

[15] To our knowledge, this technique originated by Becker and Lewis (1973) in a model studying the trade-off between quality and quantity involved in fertility choices. The linearization of non-linear budget sets was further analyzed by Edlefsen (1981) and Blomquist (1989).

the linearized program:

$$\begin{array}{rl} Max & \widetilde{U}\left(X, l, h^o, \tau\right) \\ s.c. & X + \pi_l l + \pi_{h^o} h^o = y + \pi_\tau \tau \end{array} \tag{6}$$

In particular, the demand functions appearing in the decomposition (3) are solutions to this program. The Hicksian demand functions can then be rewritten in terms of the local prices as:

$$\widetilde{l} = \widetilde{l}(\pi_l, \pi_{h^o}, \pi_\tau, \widetilde{U}) \; ; \; \widetilde{h}^o = \widetilde{h}^o(\pi_l, \pi_{h^o}, \pi_\tau, \widetilde{U}) \; ; \; \widetilde{\tau} = \widetilde{\tau}(\pi_l, \pi_{h^o}, \pi_\tau, \widetilde{U}) \tag{7}$$

In the next section we will apply this linearization technique to analyze the substitution effects arising from an adoption of the mixed compensation system.

4.2 Physicians' response to incentives: sufficient conditions

The contractual reform was introduced by the government with specific objectives in mind.

Conjecture 1 *The government expected the change in incentives induced by the switching to MR to lead to the following responses:*

- *a decrease of services provided;*
- *a decrease in the hours spent seeing patients (clinical hours of work);*
- *an increase in the hours spent on administrative and teaching duties (non clinical hours of work).*

We can now assess the reliability of this conjecture when adjustments to quality and quantity are endogenous. This requires characterizing the Hicksian demand functions of physicians. We build on the results of Lemma 2 to characterize the labor supply response to MR as a function of preferences. To this end, we first consider the trade-off between services provided and clinical hours of work (*i.e.,* gross leisure). We then consider the allocation between pure leisure and on the job leisure.

4.2.1 The leisure/services trade-off

The opportunity cost of both types of leisure is given by the same local price: $\pi_l = \pi_{h^o} = f\tau$. The sum of both, referred to as "gross leisure" and denoted $L = T - h^c$, is therefore a Hicksian aggregate.[16] This implies that the allocation of time between the two types of leisure does not vary with the price of services. We proceed in two steps, beginning with the determinants of the gross leisure/services trade-off.

Let $\pi_L = \pi_l = \pi_{h^o}$ denote the local price of gross leisure. The Hicksian demands can be rewritten as a function of this price: $\widetilde{\tau}(\pi_L, \pi_\tau, \widetilde{U})$ and $\widetilde{L}(\pi_L, \pi_\tau, \widetilde{U})$.

[16] This property relies on the assumption that the hours of work are chosen in such a way that a *per diem* is always paid under MR. See Note 13 for more details on this assumption.

Proposition 1 *The Hicksian demands for the treatment rate and clinical hours of work are both increasing in the fee paid for services if:*

- *Necessary condition:* $\eta_{h^c,\pi_\tau} = \eta_{\tau,\pi_L} < 1$;
- *Sufficient condition:* $(1-\eta_{h^c,\pi_\tau})^2 = (1-\eta_{\tau,\pi_L})^2 > \eta_{\tau,\pi_\tau}\eta_{h^c,\pi_L}$.

Proof *Compensated elasticities.* Local prices are defined as functions of practice variables. As such, they encompass the price endogeneity we stressed earlier. The non-linearity of the model implies that changing the treatment rate simultaneously affects the price paid for clinical hours of work. Taking the partial derivatives of local prices with respect to practice variables:

$$\frac{\partial \pi_L}{\partial f} = \frac{\partial \pi_l}{\partial f} = \frac{\partial \pi_{nc}}{\partial f} = \widetilde{\tau} + f\frac{\partial \widetilde{\tau}}{\partial f} \quad \text{and} \quad \frac{\partial \pi_\tau}{\partial f} = -\left[\widetilde{h}^c + f\frac{\partial \widetilde{h}^c}{\partial f}\right]$$

The compensated effect of a change in the service price on the treatment rate is

$$\frac{\partial \widetilde{\tau}}{\partial f} = \frac{\partial \widetilde{\tau}}{\partial \pi_L}\left[\tau + f.\frac{\partial \tau}{\partial f}\right] + \frac{\partial \widetilde{\tau}}{\partial \pi_\tau}\left[h^c + f.\frac{\partial \widetilde{h}^c}{\partial f}\right]$$

Using the above derivatives, and using η to denote the compensated price elasticities, gives:

$$\eta_{\tau,f} = \frac{\eta_{\tau,\pi_L} + \eta_{\tau,\pi_\tau} + \eta_{\tau,\pi_\tau}\eta_{h^c,f}}{1-\eta_{\tau,\pi_L}} \tag{8}$$

A similar line of reasoning applies to the compensated demand for gross leisure, leading to: $\eta_{L,f} = \eta_{L,\pi_L}(1+\eta_{\tau,f}) + \eta_{L,\pi_\tau}(1+\eta_{h^c,f})$. Clinical hours of work are linked to gross leisure income by the time allocation constraint. Using the fact that $L = T - h^c$, we have $\frac{\partial L}{\partial f} = -\frac{\partial h^c}{\partial f}$ and then $\frac{\partial L}{\partial f}.\frac{f}{L} = -\frac{\partial h^c}{\partial f}.\frac{f}{L}$ so that : $L.\eta_{L,f} = -h^c.\eta_{h^c,f}$. The preceding relation can be equivalently written in terms of the sensitivity of clinical hours of work to prices, giving :

$$\eta_{h^c,f_C} = \frac{\eta_{h^c,\pi_L} + \eta_{h^c,\pi_\tau} + \eta_{h^c,\pi_L}\eta_{\tau,f}}{1-\eta_{h^c,\pi_\tau}} \tag{9}$$

From the definition of compensated demands, direct local price effects are negative: $\frac{\partial \beta_j}{\partial \pi_{\beta_j}} \leq 0, \forall \beta_j = L, l, h^o, \tau$. The model with gross leisure includes only two choice variables. Assuming that consumption is a net Hicksian substitute with both types of leisure, the Euler equations imply that cross (local) price effects are positive: $\frac{\partial \widetilde{\tau}}{\partial \pi_L} \geq 0$ et $\frac{\partial \widetilde{h}^c}{\partial \pi_\tau} = -\frac{\partial \widetilde{L}}{\partial \pi_\tau} \leq 0$. Finally, the local price of the treatment rate is negative ($\pi_\tau = -fh^c$). As a result, the price elasticity of clinical hours and the treatment rate ($\eta_{\beta_j,\pi_{\beta_j}} = \frac{\partial \widetilde{\beta}_j}{\partial \pi_{\beta_j}}\frac{\pi_{\beta_j}}{\beta_j}$, $\beta_j = \{h^c, \tau\}$) are both positive.

Necessary condition. The compensated demands of clinical hours and the treatment rate are both positive if $\eta_{h^c,f} > 0$ et $\eta_{\tau,f} > 0$. Notice, the numerators of (8) and (9) are then both positive. The sensitivity of the treatment rate and clinical hours to price changes are therefore both positive only if the denominators are positive as well: $1-\eta_{\tau,\pi_L} > 0$ and $1-\eta_{h^c,\pi_\tau} > 0$. Since the Slutsky matrix is symmetric, it follows that: $\frac{\partial \widetilde{L}}{\partial \pi_\tau} = \frac{\partial \widetilde{\tau}}{\partial \pi_L}$. Using the time allocation constraint, $\frac{\partial \widetilde{L}}{\partial \pi_\tau} = -\frac{\partial \widetilde{h^c}}{\partial \pi_\tau}$, this translates into price-elasticities as : $-\frac{\partial \widetilde{h^c}}{\partial \pi_\tau}.f.\frac{h^c}{h^c} = \frac{\partial \widetilde{\tau}}{\partial \pi_L}.f.\frac{\tau}{\tau} \Leftrightarrow \eta_{h^c,\pi_\tau} = \eta_{\tau,\pi_L}$. The above necessary condition therefore reduces to: $\eta_{h^c,\pi_\tau} = \eta_{\tau,\pi_L} < 1$.

Sufficient Condition. Substituting (9) into (8) and using the symmetry of the Slutsky matrix, the elasticities can be written as:

$$\eta_{\tau,f} = \frac{\eta_{\tau,\pi_\tau}(1+\eta_{h^c,\pi_L}) + \eta_{\tau,\pi_L}(1-\eta_{\tau,\pi_L})}{(1-\eta_{\tau,\pi_L})^2 - \eta_{\tau,\pi_\tau}.\eta_{h^c,\pi_L}} \; ; \; \eta_{h^c,f} = \frac{\eta_{h^c,\pi_L}(1+\eta_{\tau,\pi_\tau}) + \eta_{h^c,\pi_\tau}(1-\eta_{h^c,\pi_\tau})}{(1-\eta_{h^c,\pi_\tau})^2 - \eta_{\tau,\pi_\tau}.\eta_{h^c,\pi_L}} \tag{10}$$

Under the necessary condition, $\eta_{h^c,\pi_\tau} = \eta_{\tau,\pi_L} < 1$, both numerators are positive. A sufficient condition for positive responses of the treatment rate and clinical hours to price changes is therefore: $(1-\eta_{h^c,\pi_\tau})^2 = (1-\eta_{\tau,\pi_L})^2 > \eta_{\tau,\pi_\tau}\eta_{h^c,\pi_L}$. ∎

The results from Proposition (1) depend in a complex way on linear cross price effects and therefore cannot be given an easy interpretation. Blomquist (1989) noted that this problem is a basic issue in utility maximization models with nonlinear budget constraints. Again, empirical analysis is required to shed light on the impact of a change in the fee paid for services on physicians' clinical hours of work and treatment rate.[17]

4.2.2 Leisure allocation

Beyond the time devoted to patients, the supply of health care also includes a variety of activities such as the management of health-care establishments and teaching. In our model, these non-clinical activities are treated as on-the-job leisure. The amount of time devoted to these tasks depends on the shape of physicians' preferences.

Proposition 2 *The relationship between leisure allocation and preferences is summarized in Table 1. Notice, in particular, that the price sensitivity of both types of leisure are negative under cases (1a), (2b), (3b), (4a), (5) and (6a,b) only.*

Table 1: Theoretical configurations of leisure allocation

Case	η_{l,π_τ}	$\eta_{h^o,\pi\tau}$	$(\eta_{h^c,\pi_L} - \eta_{\tau,\pi\tau})$	$\eta_{l,f}$	$\eta_{h^o,f}$
(1)	$+$	$-$	$+$	$+/-^a$	$-$
(2)	$+$	$-$	$-$	$-$	$+/-^b$
(3)	$-$	$+$	$+$	$-$	$+/-^b$
(4)	$-$	$+$	$-$	$+/-^a$	$-$
(5)	$-$	$-$	$+$	$-$	$-$
(6)	$-$	$-$	$-$	$+/-^a$	$+/-^b$

[a] Negative if: $\eta_{l,\pi\tau}(\eta_{h^c,f} - \eta_{\tau,f}) < (1 - \eta_{\tau,f})\eta_{l,p_x}$.
[b] Negative if: $\eta_{h^o,\pi\tau}(\eta_{h^c,f} - \eta_{\tau,f}) < (1 - \eta_{\tau,f})\eta_{l,p_x}$.

Note. The table is constructed based on the compensated elasticities (12) and (13) below. The first four columns describe the sign of the r.h.s terms in (12) and (13) ; the last two rows provide the induced signs of compensated elasticities.

Proof Given the Hicksian demand functions (7), the compensated effect of price changes on non-clinical hours of work can be written as:

$$\frac{\partial h^o}{\partial f} = \frac{\partial h^o}{\partial \pi_l}\left[\tau + f\frac{\partial \tau}{\partial f}\right] + \frac{\partial h^o}{\partial \pi_o}\left[e + f\frac{\partial \tau}{\partial f}\right] - \frac{\partial h^o}{\partial \pi_\tau}\left[h^c + f\frac{\partial h^c}{\partial f}\right]$$

[17] Following the suggestion of a referee, we derived the model using standard utility functions such as the Cobb-Douglas and the quadratic functions. Unfortunately, even in these simple cases, the model has strong nonlinearities and results are still hard to interpret. This provides another reason to use a discrete approach to estimate the model, as described below. Under a continuous approach, the likelihood function to be maximized would be very complex to estimate as particular inequality (Slutsky) conditions should hold to ensure that the function is well behaved.

Leisure allocation. The same manipulations can be applied to pure leisure. Using the definition of local prices, the allocation of leisure is described in terms of compensated elasticities by the following trade off:

$$\begin{aligned}\eta_{l,f} =&(1+\eta_{\tau,f})\left(\eta_{l,\pi_l}+\eta_{l,\pi_{h^o}}+\eta_{l,\pi_\tau}\right)+\eta_{l,\pi_\tau}\left(\eta_{h^c,f}-\eta_{\tau,f}\right)\\ \eta_{h^o,f} =&(1+\eta_{h^o,f})\left(\eta_{h^o,\pi_{h^o}}+\eta_{h^o,\pi_{h^o}}+\eta_{h^o,\pi_\tau}\right)+\eta_{h^o,\pi_\tau}\left(\eta_{h^c,f}-\eta_{\tau,f}\right)\end{aligned} \tag{11}$$

Let p_x denote the price of the consumption good. Since compensated demand functions are homogeneous of degree 0, the associated Euler equations are: $\frac{\partial\widetilde{\beta}_j}{\partial\pi_l}\pi_l+\frac{\partial\widetilde{\beta}_j}{\partial\pi_{h^o}}\pi_{h^o}+\frac{\partial\widetilde{\beta}_j}{\partial\pi_\tau}\pi_\tau=\frac{\partial\widetilde{\beta}_j}{\partial p_x}p_x$, $\beta_j\in\{l,h^o\}$. Substituting into the above system of equations, we obtain:

$$\eta_{l,f} = -\eta_{l,p_x}(1+\eta_{\tau,f})+\eta_{l,\pi_\tau}(\eta_{h^c,f}-\eta_{\tau,f}) \tag{12}$$

$$\eta_{h^o,f} = -\eta_{h^o,p_x}(1+\eta_{\tau,f})+\eta_{h^o,\pi_\tau}(\eta_{h^c,f}-\eta_{\tau,f}) \tag{13}$$

Signs. Assuming that consumption is a Hicksian substitute for both types of leisure implies: $\eta_{l,p_x}>0$ and $\eta_{h^o,p_x}>0$. Manipulating expressions (10) leads to:

$$\eta_{h^c,f}-\eta_{\tau,f}=\frac{\eta_{\tau,\pi_\tau}-\eta_{h^c,\pi_L}}{\left(1-\eta_{h^c,\pi_\tau}\right)^2-\eta_{\tau,\pi_\tau}.\eta_{h^c,\pi_L}}$$

Under the conditions given in Proposition 1, the denominator is positive. We have, moreover, established that $\eta_{L,\pi_\tau}=\eta_{l,\pi_\tau}+\eta_{h^o,\pi_\tau}<0$. It then follows that the price elasticities of non clinical hours of work and pure leisure cannot be simultaneously positive. These results are summarized in Table 1. ∎

As stated in Conjecture 1, the government expected that switching to MR would lead to decreases in both clinical hours of work and the services provided, accompanied by an increase in both types of leisure. The results summarized in Lemma 2 place this response within the context of a Slutsky decomposition derived from a model where quality and quantity are both included. With propositions 1 and 2, we can now characterize the shape of preferences that are consistent with different responses to incentives. Our findings establish that only a few configurations give rise to a behavior consistent with Conjecture (1).

To summarize: given a specific shape of preferences regarding practice (see Lemma 2 and Proposition 1),

- A voluntary switch to mixed remuneration will decrease clinical hours of work and increase the time devoted to each service;
- For cases (1a), (2b), (3b), (4a), (5) and (6a,b) in Table 1, the decrease in clinical hours of work is transmitted to both non-clinical hours of work and pure leisure;
- As a result, the time spent at work decreases;
- Under any other circumstances, the income and substitution effects are of opposite signs. Leisure allocation is therefore ambiguous.

5 Structural estimation of physicians preferences

In the previous sections we have argued that ignoring quality leads to a misspecification of the treatment effect and the social value of policies. The introduction of quality into the labor supply model creates non-linearities that complicate both the analysis of comparative statics and the results – theoretical predictions are generally ambiguous and depend on preferences. Evaluating

the effects of policies therefore requires empirical work. What is more, evaluating policies that have yet to be implemented typically requires accounting for changes in behavior on the part of economic agents, implying a role for structural econometric models (Heckman, 2000).

The same non-linearities that complicate the theoretical analysis also complicate the empirical analysis. Within the context of labor supply, the introduction of quality renders the opportunity cost of leisure endogenous – the budget constraint is generally non-linear and non-convex.

In the particular application studied here, the adoption of treatment was voluntary; namely, once MR is introduced, both FFS and MR are available to physicians. As compared with the budget constraint under FFS, MR induces two simultaneous changes. First, the discounted fee paid per service rendered causes a downwards rotation of the budget constraint. Second, the introduction of a fixed wage(*per diem*) causes an upwards shift in the budget constraint. As a result, the two budget constraints cross. The budget set constraining utility maximization is therefore non-convex. Below, we provide a discussion of the methods used for addressing this matter and a short description of the steps required for their practical implementation.

5.1 Accounting empirically for non-convex budget sets

Recovering the set of parameters governing a sample of observed choices is the general scope of a long strand of literature devoted to the structural estimation of preferences (see *e.g.*, Keane and Wolpin, 1997). Accounting for nonlinear budget constraints within these models has been an important topic of research in the labor supply literature (see Blundell and MaCurdy, 1999 for an extensive survey of this literature).

Much of the early work in this area proceeded with a piecewise linearization of the budget constraint (Burtless and Hausman, 1978; Hausman, 1980, 1985). Utility is maximized on each linear sub-segment of the budget constraint, giving the indirect utility of each local optimum. The global optimum is taken as the maximum over the local optima. A first issue with using this technique is the computational complexity of identifying each linear segment of the budget constraint. More importantly, this identification strategy has been shown to impose strong restrictions on the estimated parameters (*e.g.*, see MaCurdy, Green, and Paarsch, 1990 and MaCurdy, 1992).

A less restrictive approach is to discretize the budget constraint; see, van Soest (1995) for a general discussion of this approach and Meyer and Heim (2003) for a formal comparison with the Hausman algorithm. Applications of this approach assume that individuals maximize utility over a finite set of choices rather than along a continuous budget constraint. Within the context of the labor market, this assumption has some merit. Jobs are typically classified as part time or full time, and many people do not work at all (Zabalza, Pissarides, and Barton, 1980). Perhaps more importantly, this approach allows for considerable flexibility in the estimated parameters. The only restriction imposed is the positive marginal utility of income (van Soest, 1995). Since interior points of the budget set are excluded from estimation, this approach requires that consumption be a good.

The discrete choice approach therefore provides a flexible statistical strategy for the structural analysis of utility-maximization models with non-linear budget sets. Given the nonlinearities that are caused by the analysis of physician labor supply models with quality and quantity decisions, the evaluation of policies related to physician labour supply has much to gain from adopting this methodology.

5.2 Discretized practice model: implementation

Given data on physician choices, we want to estimate the parameters of physicians' preferences based on an optimization program such as (2). The discrete choice approach to this structural problem consists in considering only a finite number of demand choices, chosen along the budget constraint. The first step is therefore to define the choice set based on the variables appearing in the utility function. For a given set of practice variables, the compensation scheme uniquely defines the level of consumption. This step, therefore, requires a precise modeling of all the institutional aspects needed to calculate income. Structural estimation is aimed at recovering the parameters governing the observed choice among the available alternatives. The second step is, therefore, to specify the utility function that determines the observed choices. A particular discrete choice model is then determined through the distributional assumptions.

5.2.1 Choice set

Note that while the utility function is continuous, only a few points are considered along the budget constraint. The variables of interest are practice variables and consumption. For each practice variable, a finite number of choices are allowed. For example, N_c levels of clinical hours of work, N_o levels of non-clinical hours of work and N_τ levels of treatment rate. The number of levels considered is chosen by the econometrician. Computational complexity typically requires limiting the number of choices to 5 or 6, per variable. The trade-off behind this choice is clear. Computational complexity must be balanced with the variability in choice required for the identification of the model.

Utility is defined as a function not only of practice variables but also of consumption. Recall however that, conditional on a compensation scheme, consumption is uniquely defined by a given set of practice choices (Section 3.1.2). Consumption is then included in the analysis as a consequence of practice patterns rather than as a choice variable in itself. This requires precise knowledge of the real income earned as a result of different practice choices. Beyond the institutional and legal rules in force during the period of observation, data on prices and wages are then needed to estimate the model. The preferred way should of course be to collect data on actual values, hence minimizing measurement errors. An alternative solution, borrowed to the empirical literature on labor supply, is however to use estimations of wage and price functions for simulating income (see, *e.g.*, van Soest, 1995).

Given this discretization, the complete choice set of practice variables defines J alternatives among which each physician can choose. The choice set involves $dim(J) = N_c \times N_o \times N_\tau$

alternatives. A single alternative, corresponding to one particular practice possibility, is a set of values : $j = \{c_j, o_j, \tau_j\}$ respectively pointing to the ${c_j}^{th}$ level of discretized clinical hours of work, $c_j \in \{1, ..., N_c\}$, the ${o_j}^{th}$ level of discretized non-clinical hours of work, etc.

5.2.2 Distributional and functional-form specifications

We illustrate technical issues in the simpler framework of consumption/leisure trade-off. The choice inside the set of J alternatives (a choice j being in this section a level of hours of work h_j) is governed by utility maximization. We denote by $U_j = U(X_j, T - h_j)$ the deterministic part of the utility a physician derives from alternative j. This function contains the set of parameters of interest. Their precise interpretation depends upon the specification chosen by the econometrician. The large literature in demand theory devoted to the specification of utility and labor supply functions (*e.g.*, Stern, 1986) is beyond the aim of this paper. As an illustration, the most popular functional forms in discretized models are: the quadratic function in hours of work (Keane and Moffitt, 1998; Blundell, Duncan, McCrae, and Meghir, 2000), corresponding to $U(X,h) = \gamma_0 + \gamma_1 X + \gamma_2 h + \gamma_3 hX + \gamma_4 h^2 + \gamma_5 X^2$ in the traditional leisure/consumption framework, and the Translog function in leisure (van Soest, 1995; Fortin, Jacquemet, and Shearer, 2006), $U(X,(T-h)) = \gamma_0 + \gamma_1 \ln X + \gamma_2 \ln(T-h) + \gamma_3 \ln((T-h)X) + \gamma_4 (\ln h)^2 + \gamma_5 [\ln(T-h)]^2$. Whatever the specification is, observed individual heterogeneity can be incorporated through a conditional utility function. Denoting Z a vector of observed characteristics, utility can be written as $U(X, h|\, Z)$ where one – or more – of the parameters γ_m, $m = \{0, ..., 5\}$ is assumed to be a function of Z:

$$\gamma_m = \gamma_m^0 + \boldsymbol{\gamma_m^1} Z \tag{14}$$

A discrete choice model is derived from this framework by adding an unobservable random term to the utility underlying observed choices. Let ε_j be an alternative-specific random variable. The statistical model then identifies the parameters of U by assuming that physicians included in the sample make their choice of practice through the optimization of $V_j = U_j + \varepsilon_j$. The individual contribution to the likelihood of a physician choosing alternative j is then:

$$\begin{aligned} P(j) =& P[V_j \geq V_k, \quad \forall k \neq j, k \in J] \\ P(j) =& P[\varepsilon_j \geq u_k - u_j + \varepsilon_k, \quad \forall k \neq j, k \in J] \end{aligned}$$

The specification of this probability depends on the distributional assumptions. As is well known, assuming normality leads to estimate a multinomial Probit model, while a Gumbell specification gives rise to a multinomial Logit (see *e.g.*, Train, 2003 for further details on discrete choice models specification). However, note that the choice set can rapidly become very large as the levels of discretization (or, even faster, as the number of variables) are increased. Tractability has thus lead previous work to favor the multinomial Logit specification. Unlike multinomial Probits, the probability associated with Logit models can be explicitly written. This avoids the cumbersome calculations implied by the $(J-1)$ integrals involved in Probit models. The counterpart of this computational advantage is the IIA assumption imposed by the Logit specification.

In our setting, different alternatives can reflect huge variations in practice patterns ranging from, for instance, full-time research with no service at all to a high productivity full-time service delivering. Such independence between alternatives hence obviously makes little sense. What is more, the interpretation of the random terms, ε_j, must be restricted to alternative-specific arguments (typically, measurement errors) when relying on a Logit model. Both reasons make crucial to account for unobserved heterogeneity in the estimation. The most widely used way of doing so is to estimate the distribution of unobserved heterogeneity in the population through random coefficients in the utility function. Generally speaking, it implies specifying some of the coefficients γ_m according to:

$$\gamma_m \equiv \mathcal{L}(\overline{\gamma}_m, \sigma_m)$$

If observed heterogeneity is included simultaneously, the above random parameter is γ_m^0, the intercept of (14). Virtually every particular distribution can be chosen for $\mathcal{L}$, depending on the desirable correlations between coefficients and the shape that fits the best the population distribution (see Hensher and Greene, 2003, for a detailed discussion of specification issues). The above specification assumes individual unobserved heterogeneity, since parameters are invariant across alternatives. This seems well suited to labor supply applications; and mechanically induces correlations between alternatives, hence relaxing the IIA assumption. The resulting model is often called "Mixed Logit", highlighting the mixing of the Gumbell function with some heterogeneity-specific distributions. A well-known attractive feature of such a model is its ability to approximate any discrete choice model derived from random utility maximization (McFadden and Train, 2000).[18]

The cost of such a proper accounting of *(i)* correlations between alternatives and *(ii)* unobserved heterogeneity, is the reintroduction of integral computations into the estimation. This can be overcame by relying on the recent developments in simulation methods. Based on the choice probabilities derived from the above specification, the parameters governing physicians choices can then be estimated through simulated maximum likelihood.

5.2.3 Policy Evaluations

The above strategy provides a tractable framework for structural estimation with non-linearities. Given detailed data on the actual income and practice choices of physicians, application of these methods will provide estimated values of the preference parameters underlying physicians' observed choices. The change in practice (both quantity and quality dimensions), induced by any variation in the exogenous parameters (*i.e.*, notably, prices and fixed wage) can then be obtained

[18]The Logit model and its mixed version has been used in a wide range of applications in economics. One setting close to ours for technical reasons is the New Empirical Industrial Organization (starting with Berry, Levinsohn, and Pakes, 1995). Due to the focus on product penetration and competition, applications in this field tend to involve a large number of alternatives, hence the use of Logit models. The mixed Logit model was originally developed in the context of consumption (Train, McFadden, and Ben-Akiva, 1987) and transport choices (Ben-Akiva, Bolduc, and Bradley, 1993); micro-applications now range from labor economics (van Soest, 1995) to IO (Brownstone and Train, 1999).

through simulations. One simply has to resolve the model for optimal choices under the alternative policy. For example, different compensation policies can be translated into changes in the budget constraint: changes in FFS rates induce rotations in the budget constraint, the introduction of MR leads to both a rotation and a translation in the budget constraint. Simulating the response of physicians to such variations provides *ex ante* evaluations of compensation policies. For instance, Fortin, Jacquemet, and Shearer (2006) compute simulations regarding not only the free switching to MR but also the expected impact of the compulsory setting of the reform.

6 Discussion

Health-care expenses are accounting for an increasing percentage of public funds in many countries. The ability to supply health care at minimum cost is therefore becoming a policy issue of increasing importance. One tool available to governments to attain this goal is contractual design. We have argued that the evaluation of different policies must account for differences in the quality of services rendered.

With only a few exceptions (Ma, 1994; Ma and McGuire, 1997; Eggleston, 2005), health economists have paid little attention to the trade-off between quantity and quality of the supply of health services. This paper tries to fill gap by including both quantity and quality into a theoretical model of physician behavior. We extend the standard labor supply model to contexts involving multiple, multi-dimensional, work activities. We introduce quality through an objective performance measure: the treatment rate, explicitly linked to the health-care provided. This allows us to focus on the trade-off between quantity and quality in medical supply of services which gives rise to non-linearities in the budget constraint, due to price endogeneity. Adapting a linearization method borrowed to fertility literature, we show that local prices can be used to analyze contractual variation within the context of familiar classical demand theory.

Based on an unspecified utility function, our results highlight the link between the shape of preferences and the sensitivity of labor supply to incentives – the response to incentives is an empirical question. We also show that empirical evaluations of health-care policies can take account of both the quality and the quantity dimensions of the supply of health services. We provide an overview of an econometric strategy, based on discretization techniques, that allows for the structural estimation of these models.

We have presented our analysis within the context of the 1999 Quebec reform introducing a Mixed Remuneration scheme. The compensation policy under study features: a mixing between a fixed payment and a variation in the fee-for-services rate, and voluntary choice of a compensation system on the part of physicians. More generally, any change in contract will induce income and substitution effects similar to those of equation (3). Consequently, our approach applies much more generally to the evaluation of compensation policies in the presence of quality-quantity tradeoffs. An extension of our approach would be to analyze the impact of a Mixed Remuneration scheme on the health of patients. Indeed, our measure of quality is an input to health and it would be important to evaluate the effect of the reform on the consumers of health services.

Other fields of application hence include, among others, worker productivity or education.

References

Aas, I. H. M. (1995): "Incentives and financing methods," *Health Policy*, 34(3), 205–220.

Arrow, K. J. (1963): "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 53(5), 941–973.

Baker, G. P. (1992): "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 100(3), 598–614.

Becker, G. S., and H. G. Lewis (1973): "On the Interaction between the Quantity and Quality of Children," *Journal of Political Economy*, 81(2), S279–S288.

Ben-Akiva, M., D. Bolduc, and M. Bradley (1993): "Estimation of Travel Choice Models with Randomly Distributed Values of Time," *Transportation Research Record*, 1413, 88–97.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–890.

Blomquist, N. S. (1989): "Comparative Statics for Utility Maximization Models with Nonlinear Budget Constraints," *International Economic Review*, 30(2), 275–296.

Blundell, R., A. Duncan, J. McCrae, and C. Meghir (2000): "The labour market impact of the working families' tax credit," *Fiscal Studies*, 21(1), 75–104.

Blundell, R., and T. MaCurdy (1999): "Labor supply: A review of alternative approaches," in *Handbook of Labor Economics*, ed. by O. C. Ashenfelter, and D. Card, vol. 3 (1), pp. 1559–1695. North-Holland, Amsterdam.

Bolduc, D., B. Fortin, and M.-A. Fournier (1996): "The Effect of Incentive Policies on the Practice Location of Doctors: A Multinomial Probit Analysis," *Journal of Labor Economics*, 14(4), 703–732.

Brownstone, D., and K. Train (1999): "Forecasting new product penetration with flexible substitution patterns," *Journal of Econometrics*, 89(1-2), 109–129.

Burtless, G., and J. A. Hausman (1978): "The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiment," *Journal of Political Economy*, 86(6), 1103–1130.

Carlsen, F., and J. Grytten (1998): "More physicians: improved availability or induced demand?," *Health Economics*, 7(6), 495–508.

——— (2000): "Consumer satisfaction and supplier induced demand," *Journal of Health Economics*, 19(5), 731–753.

Carroll, J. (2003): "Quality Counts: So Why Not Offer Physicians Bonuses?," *Managed Care*, January.

Culyer, A. J., and R. G. Evans (1996): "Mark Pauly on welfare economics: Normative rabbits from positive hats," *Journal of Health Economics*, 15(2), 243–251.

De Jaegher, K., and M. Jegers (2000): "A model of physician behaviour with demand inducement," *Journal of Health Economics*, 19(2), 231–258.

Delattre, E., and B. Dormont (2003): "Fixed fees and physician-induced demand: A panel data study on French physicians," *Health Economics*, 12(9), 741–754.

Dranove, D. (1988): "Demand Inducement And The Physician-Patient Relationship," *Economic Inquiry*, 26(2), 281–298.

Edlefsen, L. E. (1981): "The Comparative Statics of Hedonic Price Functions and Other Nonlinear Constraints," *Econometrica*, 49(6), 1501–1520.

Eggleston, K. (2005): "Multitasking and mixed systems for provider payment," *Journal of Health Economics*, 24(1), 211–223.

Epstein, A. M. (1986): "The use of ambulatory testing in prepaid and fee-for service group practices: relation to perceived profitability," *New England Journal of Medicine*, 314, 1089–1093.

Evans, R. (1974): "Modeling the economic objectives of the physician," in *Health economics symposium, Proceedings of the First Canadian Conference 4-6 Sept.*, ed. by R. Fraser, pp. 33–46. Queen's University Industrial Relations Centre, Kingston (Ont.).

Evans, R. G. (1983): "Health Care in Canada Patterns of Funding and Regulation," *Journal of Health Politics, Policy and Law*, 8(1), 1–43.

Evans, R. G., E. M. A. Parish, and F. Sully (1973): "Medical Productivity, Scale Effects, and Demand Generation," *Canadian Journal of Economics*, 6(3), 376–393.

Feldman, R., and F. Sloan (1988): "Competition Among Physicians, Revisited," *Journal of Health Politics, Policy and Law*, 13(2), 239–262.

Ferrall, C., A. W. Gregory, and W. G. Tholl (1998): "Endogenous Work Hours and Practice Patterns of Canadian Physicians," *Canadian Journal of Economics*, 31(1), 1–27.

Fortin, B., N. Jacquemet, and B. Shearer (2006): "Compensation, Incentives and the Practice Patterns of Physicians: Theory and Evidence from Microdata," *Mimeo*.

Fuchs, V. R. (1986): "Physician-induced demand: A parable," *Journal of Health Economics*, 5(4), 367.

Gaynor, M., and P. Gertler (1995): "Moral Hazard and Risk Spreading in Partnerships," *Rand Journal of Economics*, 26(4), 591–613.

Gaynor, M., J. B. Rebitzer, and L. J. Taylor (2004): "Physician Incentives in Health Maintenance Organizations," *Journal of Political Economy*, 112(4), 915–931.

Glazer, J., and T. G. McGuire (1993): "Should physicians be permitted to 'balance bill' patients?," *Journal of Health Economics*, 12(3), 239–258.

Gosden, T., F. Forland, I. S. Kristiansen, M. Sutton, B. Leese, A. Giuffrida, M. Sergison, and L. Pedersen (2001): "Impact of payment method on behaviour of primary care physicians: a systematic review," *Journal of Health Services Research and Policy*, 6(1), 44–55.

Gosden, T., L. Pedersen, and D. Torgerson (1999): "How should we pay doctors? A systematic review of salary payments and their effect on doctor behaviour," *Quarterly Journal of Medicine*, 92(1), 47–55.

Gruber, J., and M. Owings (1996): "Physician Financial Incentives and Cesarean Section Delivery," *Rand Journal of Economics*, 27(1), 99–123.

Grytten, J., and R. Sørensen (2001): "Type of contract and supplier-induced demand for primary physicians in Norway," *Journal of Health Economics*, 20(3), 379–393.

Hausman, J. A. (1980): "The effect of wages, taxes, and fixed costs on women's labor force participation," *Journal of Public Economics*, 14(2), 161–194.

——— (1985): "The Econometrics of Nonlinear Budget Sets," *Econometrica*, 53(6), 1255–1282.

Heckman, J. J. (2000): "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *Quarterly Journal of Economics*, 115(1), 45–97.

Hemenway, D., A. Killen, S. B. Cashman, C. L. Parks, and W. J. Bicknell (1990): "Physicians' responses to financial incentives. Evidence from a for-profit ambulatory care center," *New England Journal of Medicine*, 322(15), 1059–1063.

Hensher, D., and W. Greene (2003): "The Mixed Logit model: The state of practice," *Transportation*, 30(2), 133–176.

Holmstrom, B., and P. Milgrom (1991): "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, & Organization*, 7(3), 24–52.

Hutchison, B., S. Birch, J. Hurley, J. Lomas, and F. Stratford-Devai (1996): "Do physician-payment mechanisms affect hospital utilization? A study of Health Service Organizations in Ontario," *Canadian Medical Association Journal*, 154(5), 653–661.

Keane, M., and R. Moffitt (1998): "A Structural Model of Multiple Welfare Program Participation and Labor Supply," *International Economic Review*, 39(3), 553–589.

Keane, M. P., and K. I. Wolpin (1997): "Introduction to the "JBES" Special Issue on Structural Estimation in Applied Microeconomics," *Journal of Business & Economic Statistics*, 15(2), 111–114.

Labelle, R., G. Stoddart, and T. Rice (1994a): "Editorial: Response to Pauly on a re-examination of the meaning and importance of supplier-induced demand," *Journal of Health Economics*, 13(4), 491–494.

——— (1994b): "A re-examination of the meaning and importance of supplier-induced demand," *Journal of Health Economics*, 13(3), 347–368.

Lazear, E. P. (2000): "The Power of Incentives," *American Economic Review*, 90(2), 410–414.

Levaggi, R., and L. Rochaix (2003): "Optimal payment schemes for physicians," *Portuguese Economic Journal*, 2(2), 87–107.

Ma, C.-T. A. (1994): "Health care payments systems: cost and quality incentives," *Journal of Economics & Management Strategy*, 3(1), 93–112.

Ma, C.-T. A., and T. G. McGuire (1997): "Optimal Health Insurance and Provider Payment," *American Economic Review*, 87(4), 685–704.

MaCurdy, T. (1992): "Work Disincentive Effects of Taxes: A Reexamination of Some Evidence," *American Economic Review*, 82(2), 243–249.

MaCurdy, T., D. Green, and H. Paarsch (1990): "Assessing Empirical Approaches for Analyzing Taxes and Labor Supply," *Journal of Human Resources*, 25(3), 415–490.

McFadden, D., and K. Train (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15(5), 447–470.

McGuire, T. G. (2000): "Physician Agency," in *Handbook of Health Economics*, ed. by A. J. Culyer, and J. P. Newhouse, vol. 1A, pp. 461–536. North-Holland, Amsterdam.

McGuire, T. G., and M. V. Pauly (1991): "Physician response to fee changes with multiple payers," *Journal of Health Economics*, 10(4), 385–410.

Meyer, B. D., and B. T. Heim (2003): "Structural Labor Supply Models when Budget Constraints are Nonlinear," *Working Paper*.

Nassiri, A., and L. Rochaix (2006): "Revisiting physicians' financial incentives in Quebec: a panel system approach," *Health Economics*, 15(1), 49–64.

Nguyen, N. X., and F. W. Derrick (1997): "Physician behavioral response to a Medicare price reduction," *Health Services Research*, 32(3), 283.

Paarsch, H., and B. Shearer (2000): "Piece Rates, Fixed Wages and Incentive Effects: Statistical Evidence from Payroll Records," *International Economic Review*, 41(1), 59–92.

Paarsch, H. J., and B. S. Shearer (1999): "The Response of Worker Effort to Piece Rates: Evidence from the British Columbia Tree-Planting Industry," *Journal of Human Resources*, 34(4), 643–667.

Pauly, M. V. (1994a): "Editorial: A re-examination of the meaning and importance of supplier-induced demand," *Journal of Health Economics*, 13(3), 369–372.

——— (1994b): "Reply to Roberta Labelle, Greg Stoddart and Thomas Rice," *Journal of Health Economics*, 13(4), 495–496.

Pencavel, J. (1986): "Labor supply of men: A survey," in *Handbook of Labor Economics*, ed. by O. C. Ashenfelter, and R. Layard, vol. 1, pp. 3–102. North-Holland, Amsterdam.

Rice, T. (1983): "The impact of changing Medicare reimbursement rates on physician-induced demand," *Medical Care*, 21, 803–815.

Rice, T. H., and R. J. Labelle (1989): "Do Physicians Induce Demand for Medical Services?," *Journal of Health Politics, Policy and Law*, 14(3), 587–601.

Rochaix, L. (1989): "Information asymmetry and search in the market for physicians' services," *Journal of Health Economics*, 8(1), 53–84.

——— (1993): "Financial incentives for physicians: the Quebec experience," *Health Economics*, 2(2), 163–176.

Rogerson, W. P. (1994): "Choice of treatment intensity by a no-profit hospital under prospective pricing," *Journal of Economics & Management Strategy*, 3(1), 7–51.

Rosenthal, M. B., R. Fernandopulle, H. R. Song, and B. Landon (2004): "Paying For Quality: Providers' Incentives For Quality Improvement," *Health Affairs*, 23(2), 127–141.

Schaafsma, J. (1994): "A new test for supplier-inducement and application to the Canadian market for dental care," *Journal of Health Economics*, 13(4), 407–431.

Selden, T. M. (1990): "A model of capitation," *Journal of Health Economics*, 9(4), 397–409.

Shearer, B. (2004): "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment," *Review of Economic Studies*, 71(2), 513–534.

Shekelle, P. (2003): "New contract for general practitioners," *BMJ*, 326(7387), 457–458.

Smith, P. C., and N. York (2004): "Quality Incentives: The Case Of U.K. General Practitioners," *Health Affairs*, 23(3), 112–118.

Smith, R. (2003): "The failures of two contracts," *BMJ*, 326(7399), 1097–1098.

Sørensen, R. J., and G. Jostein (1999): "Competition and supplier-induced demand in a health care system with fixed fees," *Health Economics*, 8(6), 497–508.

Stern, N. (1986): "On the Specification of Labour Supply Functions," in *Unemployment, Search and Labour Supply*, ed. by R. Blundell, and I. Walker. Cambridge University Press, Cambridge (UK).

Stiglitz, J. E. (1975): "Incentives, Risk, and Information: Notes Towards a Theory of Herarchy," *Bell Journal of Economics*, 6(2), 552–579.

Train, K. E. (2003): *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge (UK).

Train, K. E., D. L. McFadden, and M. Ben-Akiva (1987): "The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices," *RAND Journal of Economics*, 18(1), 109–123.

van Soest, A. (1995): "Structural Models of Family Labor Supply: A Discrete Choice Approach," *Journal of Human Resources*, 30(1), 63–88.

Zabalza, A., C. Pissarides, and M. Barton (1980): "Social security and the choice between full-time work, part-time work and retirement," *Journal of Public Economics*, 14(2), 245–276.